Building A World Model For A Mobile Robot Using Dynamic Semantic Constraints

Minoru Asada and Yoshiaki Shirai
Dept. of Mechanical Eng. for Computer-Controlled Machinery
Osaka University, Suita, Osaka 565, Japan
asada@ccm.osaka-u.ac.jp

Abstract

We are developing a new paradigm for a world model construction system which interprets a scene and builds a world model for a mobile robot using dynamic semantic constraints. The system represents a world model in hierarchical form from sensor-based maps to a global map with both numerical and symbolic descriptions. At the beginning of interpretation, sensory data (video and range images) are analyzed in bottom-up fashion. A range image is transformed into a height map, and analyzed for the purpose of generating a geometrical property list for both obstacle and traversable regions that is used as the initial input to the interpretation process. At each step of the scene interpretation process, the most reliable feature of an object is selected in the region property list to propagate semantic constraints on other objects close to it. Geometrical modeling for individual objects in the scene is performed, and parameters of each model are dynamically refined by the scene interpretation process. These model parameters and their interrelationships make spatial reasoning robust. Preliminary results with video and range images are shown.

1 Introduction

The development of intelligent mobile robot systems is a central problem in artificial intelligence and robotics, and has been extensively studied (ex. see many papers in the Proc. of Image Understanding Workshop [1987]). In order for these systems to accomplish various tasks such as visual navigation, obstacle avoidance, and landmark (and/or object) recognition, building and maintaining a world model from sensory data are essential problems. To construct such a world model which allows the system to represent the 3-D world and draw inferences about it, the following two major problems concerning data structures and control strategy should be addressed: (1) what kind of representation is suitable for a world model, and (2) how can the system obtain such a representation from sensory data?

Tsuji and Zheng [1987] have developed a stereo-vision-based mobile robot system that constructs a perspective map where the 3-D information obtained from stereo vision is represented in the image coordinate system. Hebert and Kanade [1986] have analyzed ERIM range images and constructed a surface property map represented in a Cartesian coordinate system viewed from top. Such a surface property map is easy to understand but does not naturally capture sensor resolution and accuracy. However, integrating several perspective maps obtained at different locations into a single perspective map seems difficult. Having both the perspective map and the 2-D map in a hierarchical representation and referring to each other when necessary is one solution for the above problem. Elfes [1987] has proposed a hierarchical representation of a sonar map in which the outputs of sonar sensors are directly mapped to a 2-D map, therefore, the difference between a sensor perspective map and an object-centered 2-D map such as surface property map is implicit. Asada [1988] has modified Elfes's system so that other sensor outputs can be represented in his hierarchy, making the relationship between the perspective map and the 2-D map explicit.

Almost all of the above systems, however, adopted data-driven analysis with bottom-up control, and do not consider a symbolic representation of a world model for task accomplishment given in a higher level description (ex. "Pick up the book on the desk behind the shelf", or "Turn right at the next intersection and stop in front of the small cabin"). In order to perform such tasks, scene interpretation and geometrical reasoning are necessary because the system has to identify each object in the scene and to draw inferences about geometrical relationships between identified objects. ACRONYM [Brooks, 1981] is a knowledge-based aerial photographs interpretation system which used general knowledge about image features and object characteristics along with specific knowledge about the objects expected in the image to guide interpretation and construct a description of the scene. Interpretation was done by an external graph matching procedure which was primarily top-down. Since the matching procedure was independent of the image data, ACRONYM could not organize its search to match the most reliable or complete features first and restrict the search for the remaining data.

In this paper, we present a new paradigm of a world model construction system which interprets a scene and builds a world model for a mobile robot by using dynamic semantic constraints. We, human beings, can easily and reliably interpret a scene using available knowledge specific to the scene. To provide a mobile robot with such a capability, not only a knowledge-base structure (frame representation [Minsky, 1975]) but also a geometrical data structure for the

world is needed. The system represents a world model in a hierarchical form from sensor-based maps to a global map with both numerical and symbolic descriptions. The result of the scene interpretation is represented as a semantic network where each node corresponding to one object is associated with the geometric model, and each arc represents geometrical relations between them. Geometrical parameters of these models described in the semantic network make spatial reasoning robust, and help with various kinds of tasks for the intelligent mobile robot systems.

2 Overview of the System

Figure 1 shows a conceptual view of the system, the hierarchical representation of a world model where various maps at different levels are included. First, sensory data (intensity and range images) are analyzed under bottom-up control. A range image is transformed to a height map, which then is segmented into traversable or obstacle regions using height information, and the intensity image is segmented by using the results of height map segmentation [Asada, 1988]. These results are stored as a sensor-based or vehicle-centered map. Second, scene interpretation using dynamic semantic constraints is applied to the results of the data-driven analysis. Domain-specific knowledge is organized as semantic constraints on objects expected in the scene and on the geometrical relations between them. They are represented as production rules, and procedures attached to them are dynamically controlled according to the contents of the current map. At each step, the most reliable feature is selected to propagate semantic constraints which effectively restrict search area for the following procedures. To realize this kind of heterarchical control, we have to evaluate the results of each procedure to find the most suitable production rule. Here, we make use of the region property list of the segmented height map. Using this list, the system can easily find the desired object feature, and then propagates semantic constraints. Third, object modeling is carried out simultaneously with scene interpretation. Each object model is represented as a frame structure with a property list, geometrical relations to other object models, and solid model parameters for artificial objects or generalized cylinder model parameters for natural objects. Whereas the solid model parameters for artificial objects are refined, and very specialized frames are selected according to the degree of interpretation (obstacle -> moving object -> automobile -> car -> Ford Probe), the generalized cylinder parameters for natural objects cannot be strictly fitted, because specialization of natural objects (obstacle -> tree -> coniferous tree -> pine tree) seems difficult and less meaningful. Thus, the world model is updated by the scene interpretation and the object modeling at each step. The final map of the world is a graph representation of a global map each node of which corresponds to a local map and each arc shows the relationship between local maps at its both ends. A local map is represented as a semantic network of object models (node) and their relationships (arc). Each node has a pointer to the corresponding object frame where property and model parameters are described. The following sections will focus on scene interpretation and object modeling.

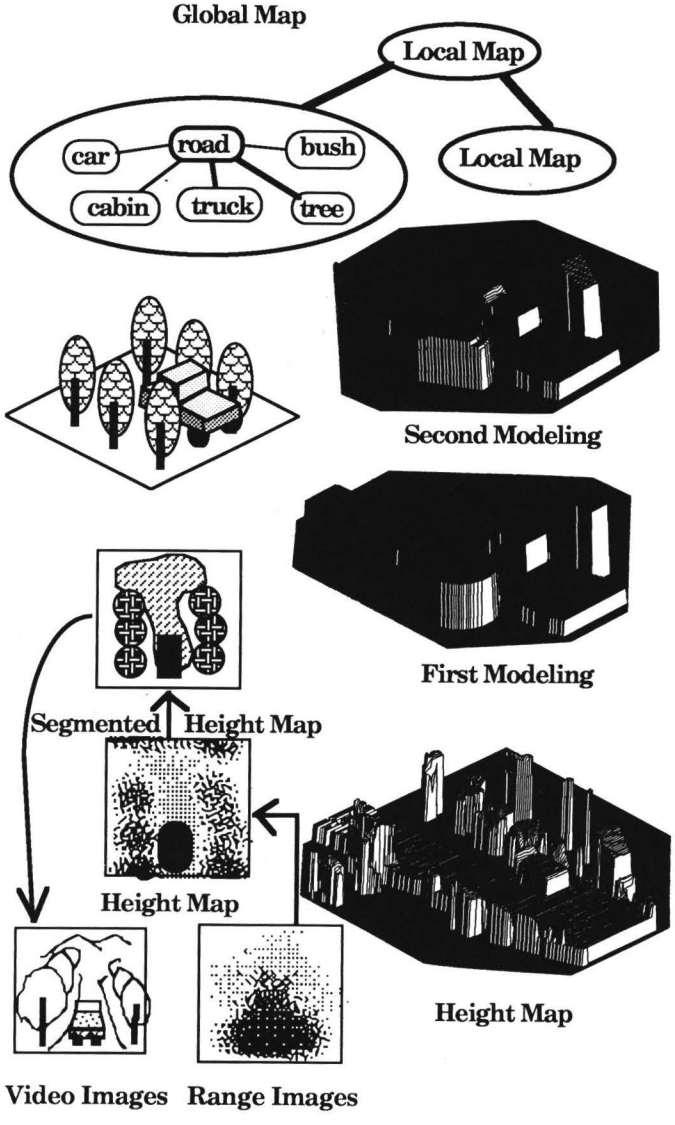

Figure 1 A conceptual view of the system; a world model construction system

3 Scene Interpretation Using Semantic Constraints

3.1 Evaluation of Bottom-Up Analysis

Our sensory data were generated by the CVL light-stripe range scanner [DeMenthon et al., 1987] developed at the Computer Vision Laboratory of the University of Maryland which was the ranging system mounted on a robot arm for simulation of a ranger-equipped vehicle. An outdoor scene was simulated by using object models (HO scale) such as trees, bushes, cabins, mail boxes, telephone poles, and cars. The input scene includes a straight road, a T-type intersection, two cabins, one truck, two cars, a mailbox, a stop sign at the intersection, trees and bushes (see Figure 4(c)). A range image (sensor map, ex. see Figure 2(a)) was transformed to a height map (Figure 2(b)) with respect to a mobile robot. In the range image (height map), the darker points are closer to the viewer (higher with respect to the assumed ground plane), and the brighter points are farther

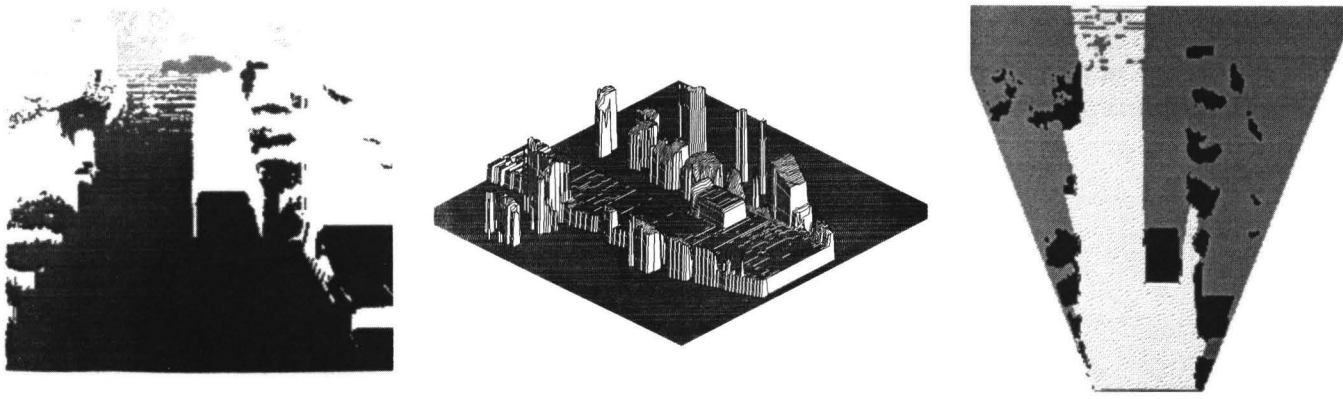

Figure 2 The results of bottom-up analysis; (a) range image [left], (b) height map [center], and (c) segmentation of (b) [right]

(lower) from it. In the white regions, range (height) information is not available due to inadequate reflection or occlusion. The height map was segmented into four categories (see Figure 2(c); unexplored (white), occluded (dark gray), traversable (light gray) or obstacle (black) regions) using height information for obstacle detection and path planning. One drawback of the height map - recovery of vertical planes is not possible - was overcome by the utilization of multiple height maps which include the maximum and minimum heights of each point, and the number of points in the range image mapped to one point in the height map. The multiple height map was useful not only for finding vertical planes in the height map, but also for segmentation of the intensity image (Ex., see bushes on the left side in Figure 4(c)). See [Asada, 1988] for more detail. The results of height map segmentation are transformed into a region property list where labels (traversable or obstacle) and geometrical properties (location, area, mean height, mean slope, mean curvature, variances of slope and curvature) are described for each region for subsequent scene interpretation.

3.2 Knowledge Representation and Road Scene Interpretation

The scene interpretation system consists of three parts; a knowledge-base, a working memory, and an inference engine. The knowledge base has three kinds of knowledge; the top level is domain-specific knowledge for road scene interpretation, the second level is general knowledge of geometrical descriptions (ex., "parallel lines" and "smooth curves"), and geometrical relations ("close", "along", "on" and so on), and the third level is general knowledge about image processing (ex., edge finding, line fitting and so on). Production rules for identifying each object in the scene (top level), for reasoning geometrical relationships (second level), and for handling image processing software (third level) are organized for utilization of the above knowledge base. The region property list obtained from bottom-up analysis of sensory data is stored in the working memory. The principle of the control strategy included in the inference engine whose is to identify object labels as more specific one (an instance) in the working memory (downward in Figure 3). The following sections focus on the representation of domain-specific knowledge and preliminary results of scene interpretation.

3.2.1 Domain-specific knowledge

In order to identify obstacle or traversable regions as more specific objects such as trucks, cars, cabins, and trees, the system needs domain-specific knowledge. Here, we organize the domain-specific knowledge as a network of frame structures of object models expected in the scene (see Figure 3). In each frame, semantic constraints on the object model and on the relationship to other object models are described. The followings are examples of road, automobile, and bush frames.

(Road Frame
    (IS-A traversable-region)
    (CONSIST-OF a-pair-of-road-boundaries center-part)
    (PROPERTY constant-width smooth-curve)
    (RELATION-TO-OTHER-OBJECTS
        (ON-THE-ROAD automobile)
        (ALONG-THE-ROAD bush building automobile other-objects)))

(Automobile Frame
    (IS-A artificial-object)
    (PROPERTY movable)
    (SIZE
        (WIDTH min-width max-width)
        (LENGTH min-length max-length)
        (HEIGHT min-height max-height))

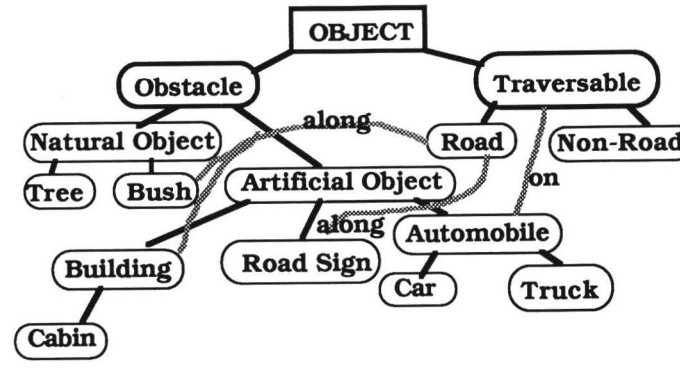

Figure 3 A network of frame structure for domain-specific knowledge

(RELATION-TO-OTHER-OBJECTS
(MOVING-ON traversable-region)
(STOP-IN-FRONT-OF stop-sign obstacle)
(STATIONARY-AT parking-area)))

(Bush Frame
(IS-A natural-object)
(PROPERTY
(LOWER-THAN tree)
(RELATION-TO-OTHER-OBJECTS
(ALONG road-boundary)))

In the "Bush Frame", other properties such as high mean curvature, large variances of slope and curvature should be described since a bush has these properties. They are, however, included in its upper class, the "natural object" frame, and "Bush Frame" inherits them from it. The following items are examples of rules used to identify each object in the scene on the top level.

.

#301 : If there is a pair of parallel smooth curves in the traversable region,
then match the "Road Frame".
#302 : If there is a moving object on the road,
then match the "Automobile Frame".
#303 : If there is an object near the road,
then match the individual frames of bush, building, parked automobile, or other object.

.

#401 : If an object has high curvature, large variances of slope and curvature
then match the "Natural-Object Frame".

3.2.2 Preliminary Results

First of all, the region property list as the result of height map segmentation is stored in the working memory. The larger region close to the viewer is selected because it is the most reliable feature in the list. In a road scene, a traversable region would be the largest in the working memory and is selected as the first object feature. If there are plural candidates for the first feature, the system chooses an arbitrary one and keeps the remaining ones for later interpretation, expecting that the semantic constraints from the first one propagate to other candidates. In our experiments, a traversable region is the largest in the working memory and is selected as the first object feature. In order to interpret a road scene, a road region is a key feature for propagating semantic constraints on objects expected in the scene. Since road regions should be included in the traversable region, the system tries to detect the road boundaries in the traversable region. In the height map, finding road boundaries seems difficult. Therefore, the system maps the traversable region from the height map to the intensity image in order to find road boundaries using the intensity information (see Figure 4(a)). A road edge finder is applied to that part of the intensity image which corresponds to the traversable region, and detected road region boundaries are mapped onto the height map for later processing. Figures 5(b) and (c) indicate the results of road boundary detection. Although there are some strong edges near the right road boundary caused by the shadow of the cabin on the right side, only short segments remain because they have various kinds of orientations along the shadow line due to texture patterns on the road surface. Figure 4(b) shows the final result of road region detection, and Figure 4(c) indicates the corresponding road region on the height map.

The next step is to find and identify objects near the obtained road region. Candidates for them are automobiles (on the road), bushes, trees, buildings (along the road), and so on. When one of them becomes evident in the height map and/or the intensity image, the corresponding object frame is called to identify it. The first candidate is an obstacle (truck) touching the road region. There are two possible interpretations for the object; one is an object bounding the road, and the other is an obstacle on the road. The probability of the former is not so high because the road width changes suddenly, therefore, the later one is kept by the system. Since

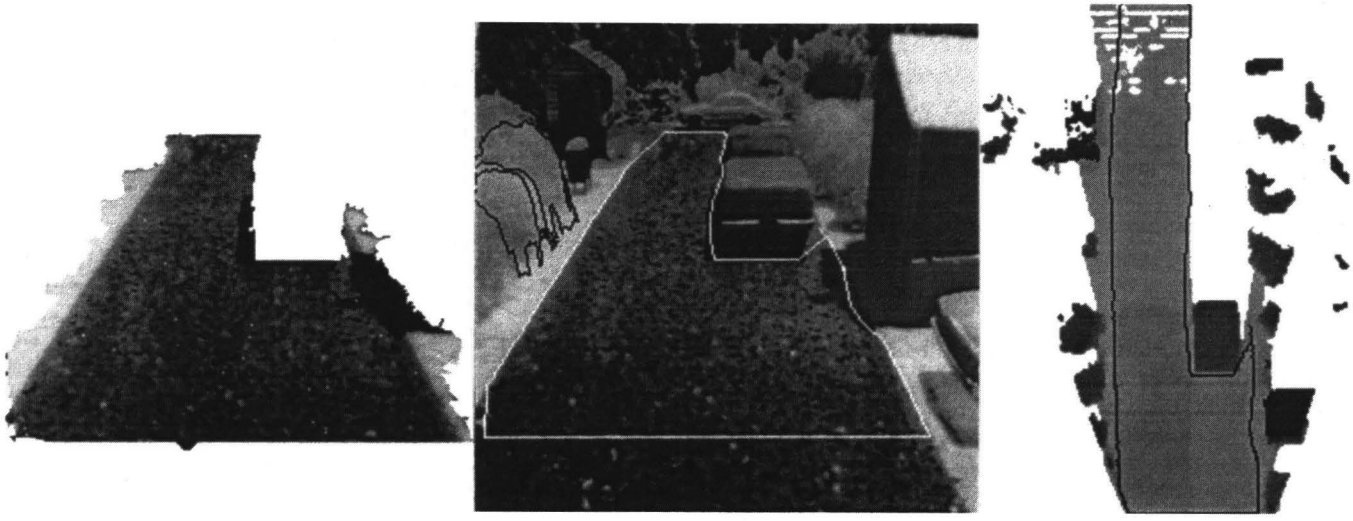

Figure 4 Road boundary detection; (a) traversable region in intensity image [left], (b) road region in intensity image [center], and (c) road region on height map [right]

the probability of the object to be an automobile is very high if an obstacle is on the road, this hypothesis is verified by a follow-up observation. There are two objects selected as being near the road region. One is an object (bush) along the left road boundary, and another one (cabin) near the right road boundary. Candidates of the objects near the road boundary are bushes, trees, buildings, parked cars and so on according to rule #303. The left side object is hypothesized as a natural object (bush or tree) because of its high mean curvature, large variances of slope and curvature (rule #401). The shape of the boundary of this region on the intensity image has many curvilinear segments (see Figure 4(b)), and this make its interpretation as a natural object more likely. The right side object is hypothesized as an artificial object due to its planar surface (from the PROPERTY slot in the "Artificial-object Frame"). Thus, the most reliable object feature is selected in each step in order to propagate semantic constraints to other objects in the working memory, and interpretation of intensity image and height map is efficiently performed.

4 Geometrical Modeling with Solid or GC Model

The final goal of our research is not only to represent a world model by symbolic descriptions, but also to provide a geometrical model associated with it so that the system can easily reason about the 3-D world. To realize such a world model, both object modeling and scene interpretation processes should be carried out simultaneously, because they cannot make use of feedback from the intermediate results from each other. if the object modeling was performed after scene interpretation.

4.1 Solid Modeling For Artificial Objects

Object modeling should satisfy the following requirements: (1) the number of parameters should be small (efficient representation), and (2) redundancy should be low (correct representation). We adopt the solid model with planar surfaces for artificial objects to satisfy the above requirements since artificial objects such as houses, road signs, cars, and trucks have many planar surfaces. The purpose of object modeling is to provide approximate location, size, and shape for geometrical reasoning which helps the scene interpretation. To reflect the intermediate results of scene interpretation, the system has three levels of approximation of solids for each region according to object specialization.

On the first approximation level (artificial object), a box model is used to represent the location and size of the object. The box model consists of four vertical walls and one horizontal top surface, all of which are touching with some parts of the object; in other words, the model is the bounding box of the object on the height map. The next level (instance of artificial objects such as cabin, car, truck, or road sign) is a polyhedron model any surface of which is touching an edge or surface of the object. The final level is a composite model with boxes, wedges, and planar patches for more specialized object such as Ford Probe, an instance of car. Although vertical planes can be detected from the multiple height map [6], close examination of the range images is needed to determine the size of each part correctly, for example, overhang of the roof.

4.2 Generalized Cylinder Modeling For Natural Objects

The only purpose of natural object modeling is to represent location and size for each region because, unlike artificial objects, determining accurate geometrical parameters for natural objects seems difficult and less meaningful for mobile robot systems. If a natural object is selected as a landmark, outstanding feature(s) of color and/or size such as deep red, very large, and/or very tall is necessary rather than fine structure of the object. Here, we adopt the Straight Homogeneous Generalized Cylinder (SHGC in short) model for natural objects. Bushes and trees grow based on their trunk whose orientation is generally vertical due to gravity. Thus, we select the vertical direction as axis orientation, and smoothly change the radius of the circle of the cross-section in order to cover the object. Splitting a SHGC into plural small SHGCs can be performed if necessary.

4.3 Preliminary Results

Figure 5(a) shows the first approximations of four objects; road (planar patch), automobile on the road (box), natural object on the left side (right cylinder), and artificial object on the right side (box). To specialize the object hypothesized as an automobile, the system picks the "car" and "truck" frames from the SIZE slot of them (the maximum height is determined from the height map although the length (or width) is not determinable due to self-occlusion). The second approximation is carried out to specialize this object. The planar region as a supporting plane of the truck is found, which increases the probability of "truck". From the fitting of planes to the object on the right side, the "cabin" frame is selected, and parameters of the plane, height and slope, for the part of the roof are refined. The right cylinder for the natural object is also reshaped by the change of the radius of the cross-section circle. Figure 5(b) illustrates the above processes. Thus, scene interpretation by specialization of region labels in the working memory is carried out, and the parameters of geometrical models are refined.

5 Discussion

We are developing a world model construction system, and the interpretation of video and range images for a mobile robot with dynamic semantic constraints was described. At each step, the results of the scene interpretation is reflected onto geometrical modeling of each object in the scene, and the refined model makes scene interpretation efficient and robust. Use of new sensory data observed at different locations would be helpful for verification or correction of the scene interpretation (especially, for moving objects). Before utilizing new sensory data, we have to solve the correspondence problem. Although finding the correspondence between two images is generally difficult, the system can find the correspondence of two height maps efficiently using the region property list. Scene interpretation can be verified and updated by using multiple views, and the geometrical parameters of each object would become more determinable as their interpretations become more specialized. We have dis-

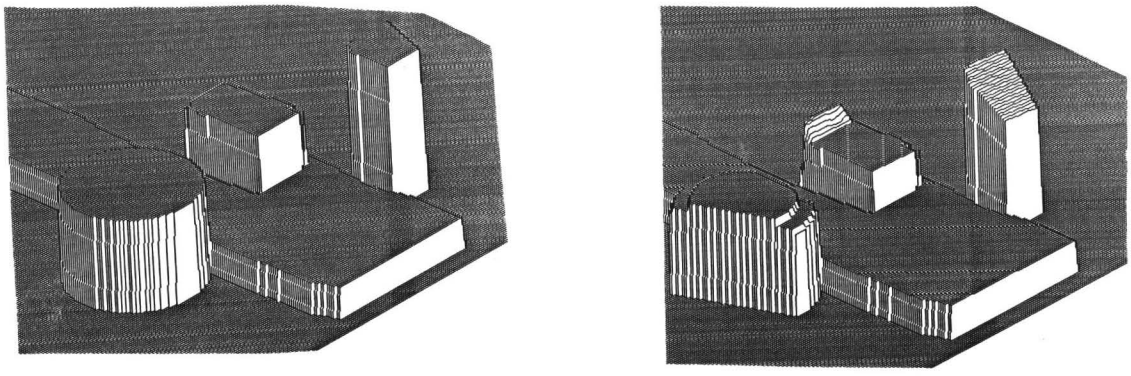

Figure 5 Geometrical parameter refinement; (a) first approximation with box and right cylinder [left], and (b) second approximation with polyhedron and SHGC [right]

cussed this problem in [Asada et al., 1989]. The following problems should be addressed in future work:

Representation of the uncertainty of scene interpretation; in the current system, certainty of the scene interpretation is empirically determined, and the rule for fusing different views is very simple (average). Exact definition of certainty for scene interpretation is needed, and fusing rules for scene interpretations at different views, which are consistent with the defined certainties, should be developed.

Various kinds of low-level signal processing are needed and should be robust, and their limitations should be described exactly. Otherwise, scene interpretation would be unstable, unreliable, and infeasible. The number of available programs for signal-level processing is limited in the current system.

Acknowledgement

The first author wishes to thank Mr. Daniel DeMenthon for providing range images at the University of Maryland, Computer Vision Laboratory. He also wishes to thank Prof. Saburo Tsuji and Mr. Michael Hild at Osaka University, Japan, for constructive discussions. Ryu Asada, a son of the first author contributed to the research by assembling and painting the landscape models on the simulation board.

References

[Proc. of Image Understanding Workshop, 1987] In *Proceeding of Image Understanding Workshop,* February 1987.

[Tsuji and Zheng, 1987] Saburo Tsuji and Jiang Yu Zheng. Visual path planning by a mobile robot. In *Proceeding of the Tenth International Joint Conference on Artificial Intelligence,* pages 1127-1130, Milan, Italy, August 1987. International Joint Committee on Artificial Intelligence.

[Hebert and Kanade, 1986] Martial Hebert and Takeo Kanade. Outdoor scene analysis using range data. In *Proceeding of the IEEE International Conference on Robotics and Automation,* pages 1426-1432, 1986.

[Elfes, 1987] Alberto Elfes. A sonar-based real world mapping and navigation. *IEEE Journal of Robotics and Automation,* vol.RA-3, pages 249-265, 1987.

[Asada, 1988] Minoru Asada. Building 3-D world model for a mobile robot from sensory data. In *Proceeding of IEEE International Conference on Robotics and Automation,* pages 918-923, 1988.

[Brooks, 1981] Rodney A. Brooks. Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence,* 17:285-348, 1981.

[Minsky, 1975] Marvin Minsky. *A framework for representing knowledge.* In: The Psychology of Computer Vision, P. H. Winston (ed.), McGraw-Hill, New York, 1975.

[DeMenthon et al., 1987] Daniel DeMenthon, Tharakesh Siddalingaiah, and Larry S. Davis. Production of dense range images with the CVL light-stripe range scanner. Center for Automation Research Technical Report CAR-TR-337, University of Maryland, 1987.

[Asada et al., 1989] Minoru Asada, Eiji Ikeda, and Yoshiaki Shirai. Interpretation and integration of height maps from a range image sequence. In *Proceeding of IEEE Workshop on Intelligent Robot and Systems,* 1989 (to appear).